Data Mining: Lecture 2

agenda

Major Issues in Data Mining

Major Algorithms in Data Mining

Major Venues of Data Mining

MAJOR ISSUES IN DATA MINING

Are All the "Discovered" Patterns Interesting?

- Data mining may generate thousands of patterns: Not all of them are interesting
  - Suggested approach: Human-centered, query-based, focused mining
- **Interestingness measures**
  - A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
- **Objective vs. subjective interestingness measures**
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.

Find All and Only Interesting Patterns?

- Find all the interesting patterns: Completeness
  - Can a data mining system find all the interesting patterns? Do we need to find all of the interesting patterns?
  - Heuristic vs. exhaustive search
  - Association vs. classification vs. clustering
- Search for only interesting patterns: An optimization problem
  - Can a data mining system find only the interesting patterns?
  - Approaches
    - First generate all the patterns and then filter out the uninteresting ones
    - Generate only the interesting patterns—mining query optimization

Other Pattern Mining Issues

- Precise patterns vs. approximate patterns
  - Association and correlation mining: possible find sets of precise patterns
    - But approximate patterns can be more compact and sufficient
    - How to find high quality approximate patterns??
- Constrained vs. non-constrained patterns
  - Why constraint-based mining?
  - What are the possible constraints? How to push constraints into the mining process?

Why Not Traditional Data Analysis?

- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

Multi-Dimensional View of Data Mining

- **Data to be mined**
  - Relational, data warehouse, transactional, stream, object-oriented/relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW
- **Knowledge to be mined**
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

Data Mining: Classification Schemes

- Different views lead to different classifications
  - Data view: Kinds of data to be mined
  - Knowledge view: Kinds of knowledge to be discovered
  - Method view: Kinds of techniques utilized
  - Application view: Kinds of applications adapted

Data Mining: On What Kinds of Data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

Relational Database

- Interrelated data
- Software programs
- to manage & access
- Data
- Data access
  - Concurrent
  - Shared
  - Distributed
- Based on ER model
- Contains a unique key
- SQL Vs. Data mining
  - SQL: Look for customers or sales in a month
  - Data mining: determine credit risk of customers

customer

cust_ID	*name*	*address*	*age*	*income*	*credit_info*	*category*
C1	Smith, Sandy	1223 Lake Ave., Chicago, IL	31	$78000	1	3
. . .	. . .	. . .	. . .	. . .	. . .	. . .

item

item_ID	*name*	*brand*	*category*	*type*	*price*	*place_made*	*supplier*	*c*
I3	hi-res-TV	Toshiba	high resolution	TV	$988.00	Japan	NikoX	$
I8	Laptop	Dell	laptop	computer	$1369.00	USA	Dell	$
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. .

employee

empl_ID	*name*	*category*	*group*	*salary*	*commission*
E55	Jones, Jane	home entertainment	manager	$118,000	2%
. . .	. . .	. . .	. . .	. . .	. . .

branch

branch_ID	*name*	*address*
B1	City Square	396 Michigan Ave., Chicago, IL
. . .	. . .	. . .

purchases

trans_ID	*cust_ID*	*empl_ID*	*date*	*time*	*method_paid*	*amount*
T100	C1	E55	03/21/2005	15:45	Visa	$1357.00
. . .	. . .	. . .	. . .	. . .	. . .	. . .

items_sold

trans_ID	*item_ID*	*qty*
T100	I3	1
T100	I8	2
. . .	. . .	. . .

works_at

empl_ID	*branch_ID*
E55	B1
. . .	. . .

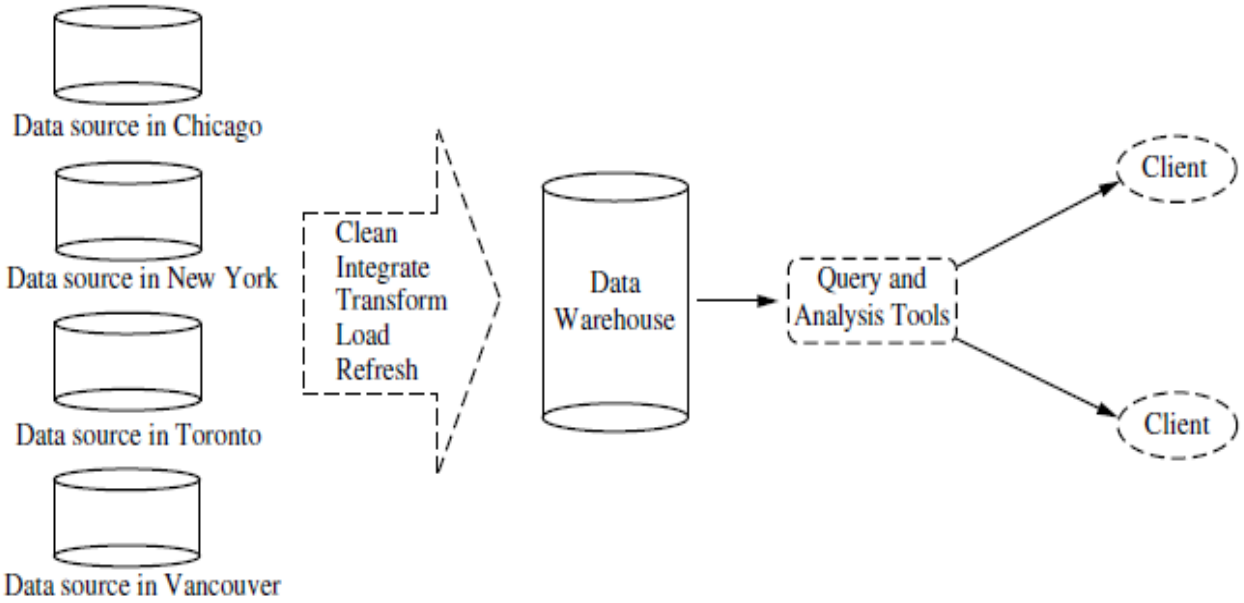
Transactional Database

- File where each record represents a transaction
- Normal queries
  - Items bought by Zafar Iqbal
  - Transactions for a certain item such as cigarettes etc.
- Data mining can
  - Find what items are sold together (market basket)
  - What items are more frequently sold

trans_ID	*list of item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
...	...

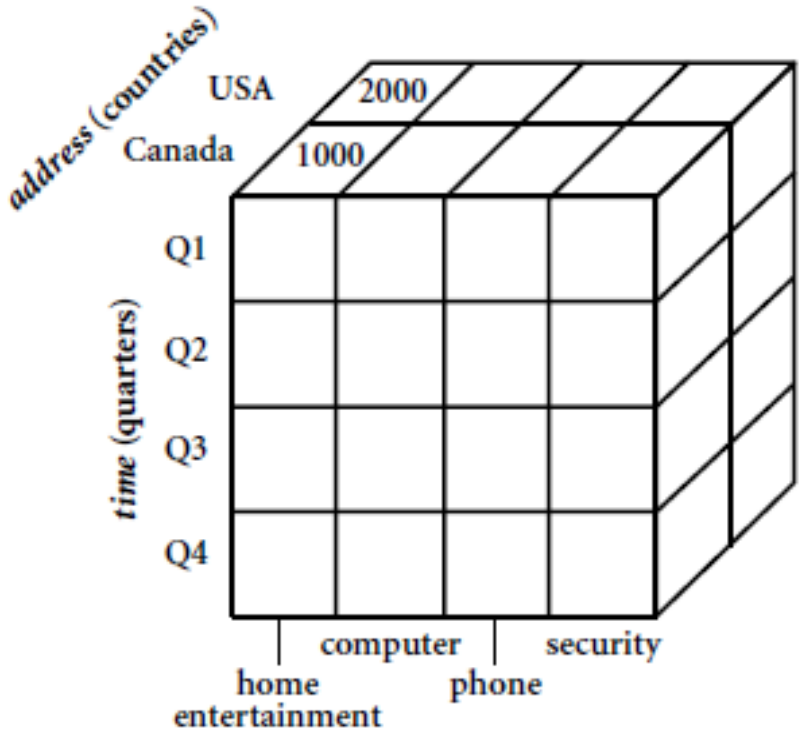
Datawarehouse

- Summary of data organized around major subjects
  - Involves data cleaning, integration, transformation, loading and periodic refreshing
- Multi-dimensional database structure
  - Each dimension corresponds to an attribute



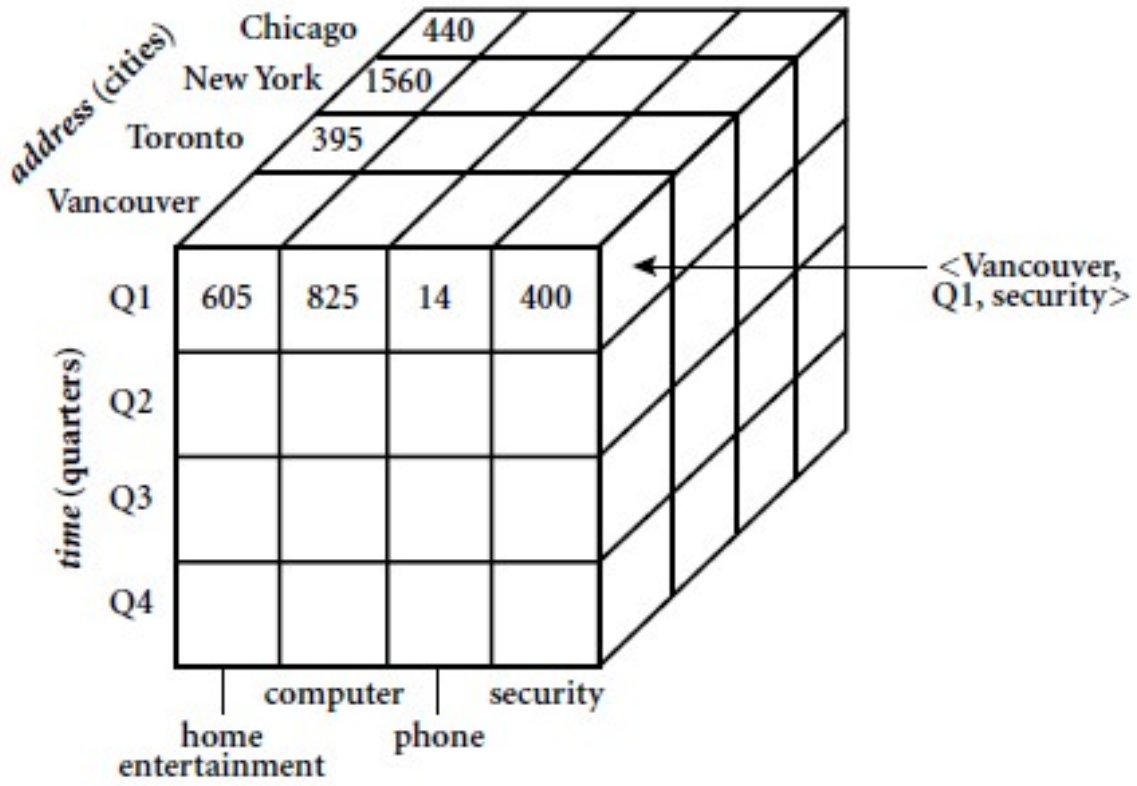
Data mart vs. Datawarehouse: Department wide vs. enterprise wide

Data-cubes Level 1



Lets “drill down” on countries!

Data-cubes Level 2



Lets ‘drill down’ on time!

Advanced Data and Information Systems

- Object Relational Databases
- Temporal, Sequence and Time-series Databases
  - Examples: data from stock exchange, inventory control and observation of natural phenomena
  - Data mining to unravel the change in trends
- Spatial and Spatiotemporal Databases
  - Uncover patterns pertaining fields, gardens or houses
- Text and Multimedia Databases
- Heterogeneous and Legacy Databases
  - Information exchange is the main issue which may be resolved using Data mining to generalize data in to higher conceptual levels
- Data Streams
  - Scientific and engineering data
  - Data mining-constantly evaluate incoming streams for patterns and dynamic changes
- The World Wide Web
  - Web mining

Other Major Issues in Data Mining

- Mining methodology
  - Mining different kinds of knowledge from diverse data types, e.g., bio, stream, Web
  - Performance: efficiency, effectiveness, and scalability
  - Pattern evaluation: the interestingness problem
  - Incorporation of background knowledge
  - Handling noise and incomplete data
  - Parallel, distributed and incremental mining methods
  - Integration of the discovered knowledge with existing one: knowledge fusion
- User interaction
  - Data mining query languages and ad-hoc mining
  - Expression and visualization of data mining results
  - Interactive mining of knowledge at multiple levels of abstraction
- Applications and social impacts
  - Domain-specific data mining & invisible data mining
  - Protection of data security, integrity, and privacy

MAJOR ALGORITHMS IN DATA MINING

Data Mining Functionalities

- Data mining tasks can be classified in to two categories:
  - Descriptive: Characterize the general properties of data
  - Predictive: Inferences on current data in order to make predictions
- A measure of certainty may also be associated with each pattern

Data Mining Functionalities

- Multidimensional concept description: Characterization and discrimination
  - Characterization: Generalize or summarize the target class or class under study based upon features, and contrast data characteristics, e.g., dry vs. wet regions
  - Discrimination: Is comparing a target class with a set of contrasting classes
- Classification and prediction
  - Construct models (functions) that describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)

Classification

(a)

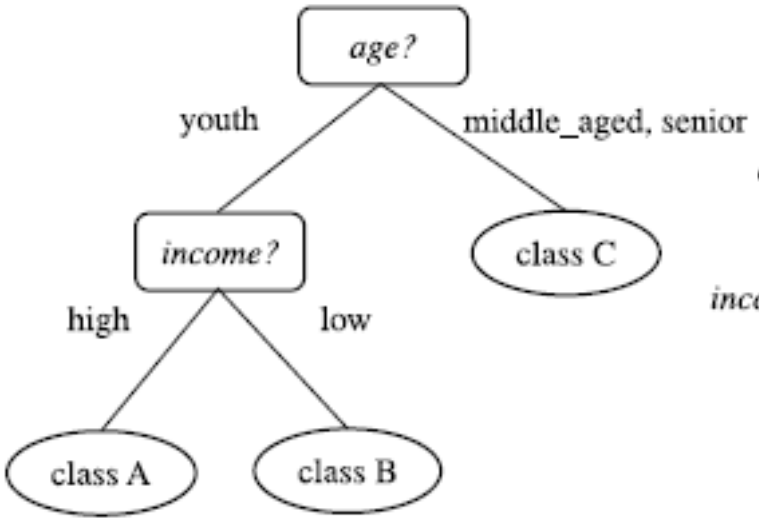
age(X, "youth") AND income(X, "high") ⟶ class(X, "A")

age(X, "youth") AND income(X, "low") ⟶ class(X, "B")

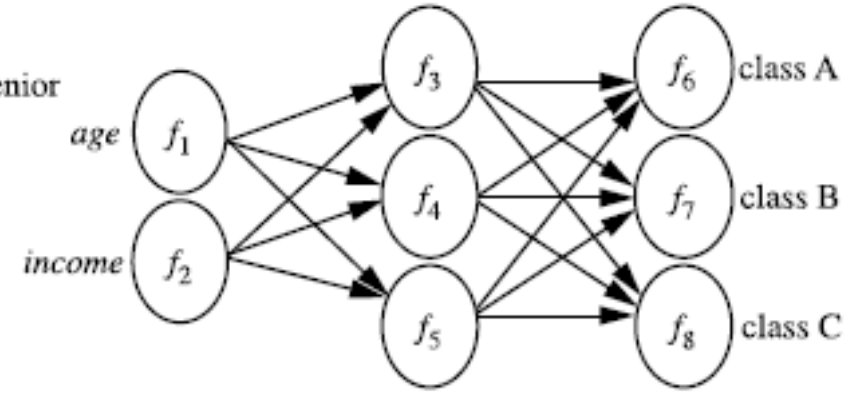
age(X, "middle_aged") ⟶ class(X, "C")

age(X, "senior") ⟶ class(X, "C")

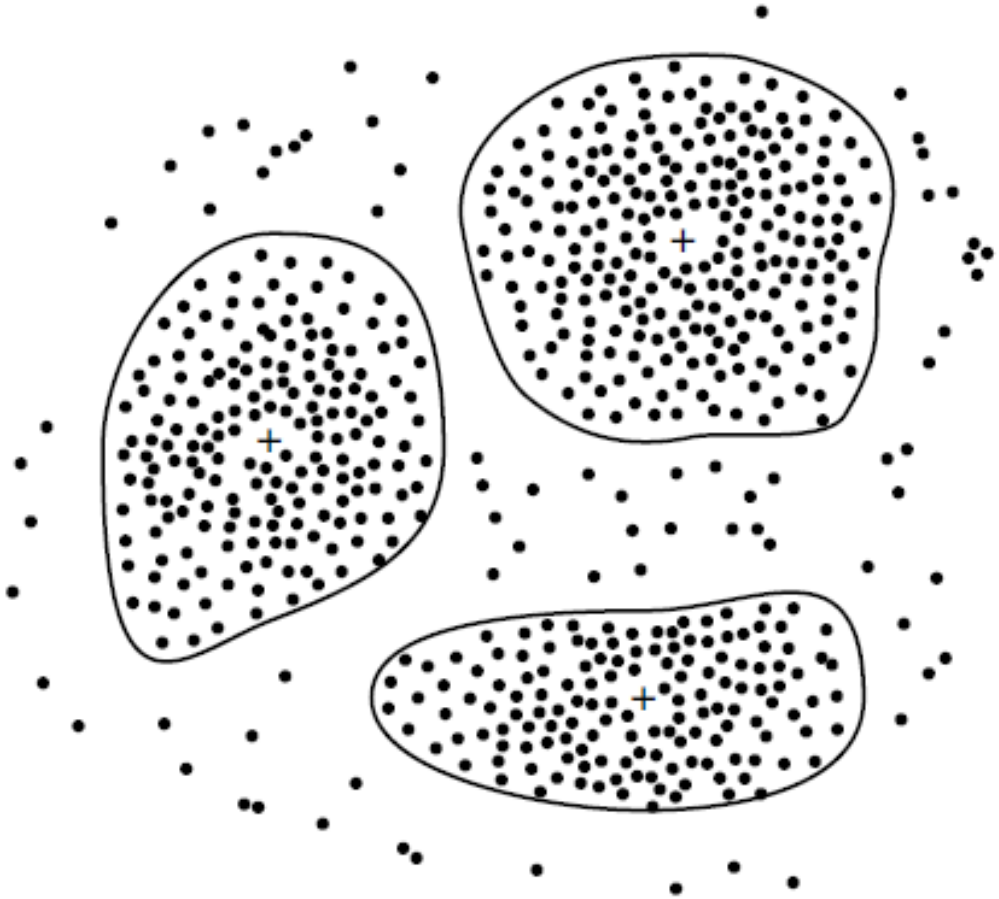
(b)



(c)



Clustering



Data Mining Functionalities

- Cluster analysis
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Maximizing intra-class similarity & minimizing interclass similarity
- Outlier analysis
  - Outlier: Data object that does not comply with the general behavior of the data
  - Noise or exception? Useful in fraud detection, rare events analysis
- Trend and evolution analysis
  - Trend and deviation: e.g., regression analysis
  - Sequential pattern mining: e.g., digital camera large SD memory
  - Periodicity analysis
  - Similarity-based analysis

Top-10 Most Popular DM Algorithms:
18 Identified Candidates (I)

- **Classification**
  - #1. C4.5: Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann., 1993.
  - #2. CART: L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth, 1984.
  - #3. K Nearest Neighbours (kNN): Hastie, T. and Tibshirani, R. 1996. Discriminant Adaptive Nearest Neighbor Classification. TPAMI. 18(6)
  - #4. Naive Bayes Hand, D.J., Yu, K., 2001. Idiot's Bayes: Not So Stupid After All? Internat. Statist. Rev. 69, 385-398.
- **Statistical Learning**
  - #5. SVM: Vapnik, V. N. 1995. The Nature of Statistical Learning Theory. Springer-Verlag.
  - #6. EM: McLachlan, G. and Peel, D. (2000). Finite Mixture Models. J. Wiley, New York. Association Analysis
  - #7. Apriori: Rakesh Agrawal and Ramakrishnan

The 18 Identified Candidates (II)

- **Link Mining**
  - #9. PageRank: Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In WWW-7, 1998.
  - #10. HITS: Kleinberg, J. M. 1998. Authoritative sources in a hyperlinked environment. SODA, 1998.
- **Clustering**
  - #11. K-Means: MacQueen, J. B., Some methods for classification and analysis of multivariate observations, in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, 1967.
  - #12. BIRCH: Zhang, T., Ramakrishnan, R., and Livny, M. 1996. BIRCH: an efficient data clustering method for very large databases. In SIGMOD '96.
- **Bagging and Boosting**
  - #13. AdaBoost: Freund, Y. and Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 1 (Aug. 1997), 119-139.

The 18 Identified Candidates (III)

- **Sequential Patterns**
  - #14. GSP: Srikant, R. and Agrawal, R. 1996. Mining Sequential Patterns: Generalizations and Performance Improvements. In Proceedings of the 5th International Conference on Extending Database Technology, 1996.
  - #15. PrefixSpan: J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and M-C. Hsu. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In ICDE '01.
- **Integrated Mining**
  - #16. CBA: Liu, B., Hsu, W. and Ma, Y. M. Integrating classification and association rule mining. KDD-98.
- **Rough Sets**
  - #17. Finding reduct: Zdzislaw Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Norwell, MA, 1992
- **Graph Mining**
  - #18. gSpan: Yan, X. and Han, J. 2002. gSpan: Graph-Based Substructure Pattern Mining. In ICDM '02.

Top-10 Algorithm Finally Selected at ICDM'06

- **#1: C4.5 (61 votes)**
- **#2: K-Means (60 votes)**
- **#3: SVM (58 votes)**
- **#4: Apriori (52 votes)**
- **#5: EM (48 votes)**
- **#6: PageRank (46 votes)**
- **#7: AdaBoost (45 votes)**
- **#7: kNN (45 votes)**
- **#7: Naive Bayes (45 votes)**
- **#10: CART (34 votes)**

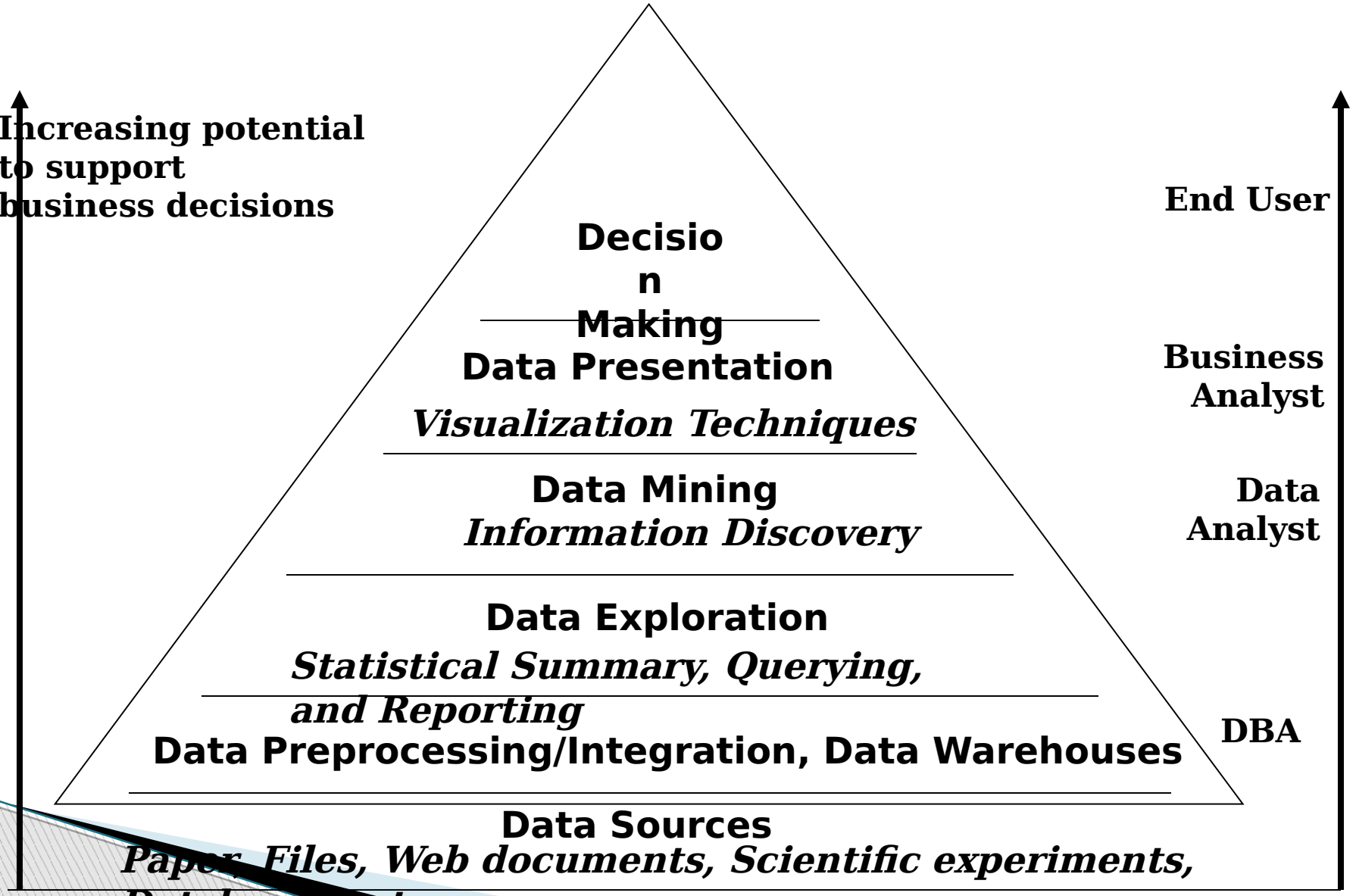
A Brief History of Data Mining Society

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.
- ACM Transactions on KDD starting in 2007

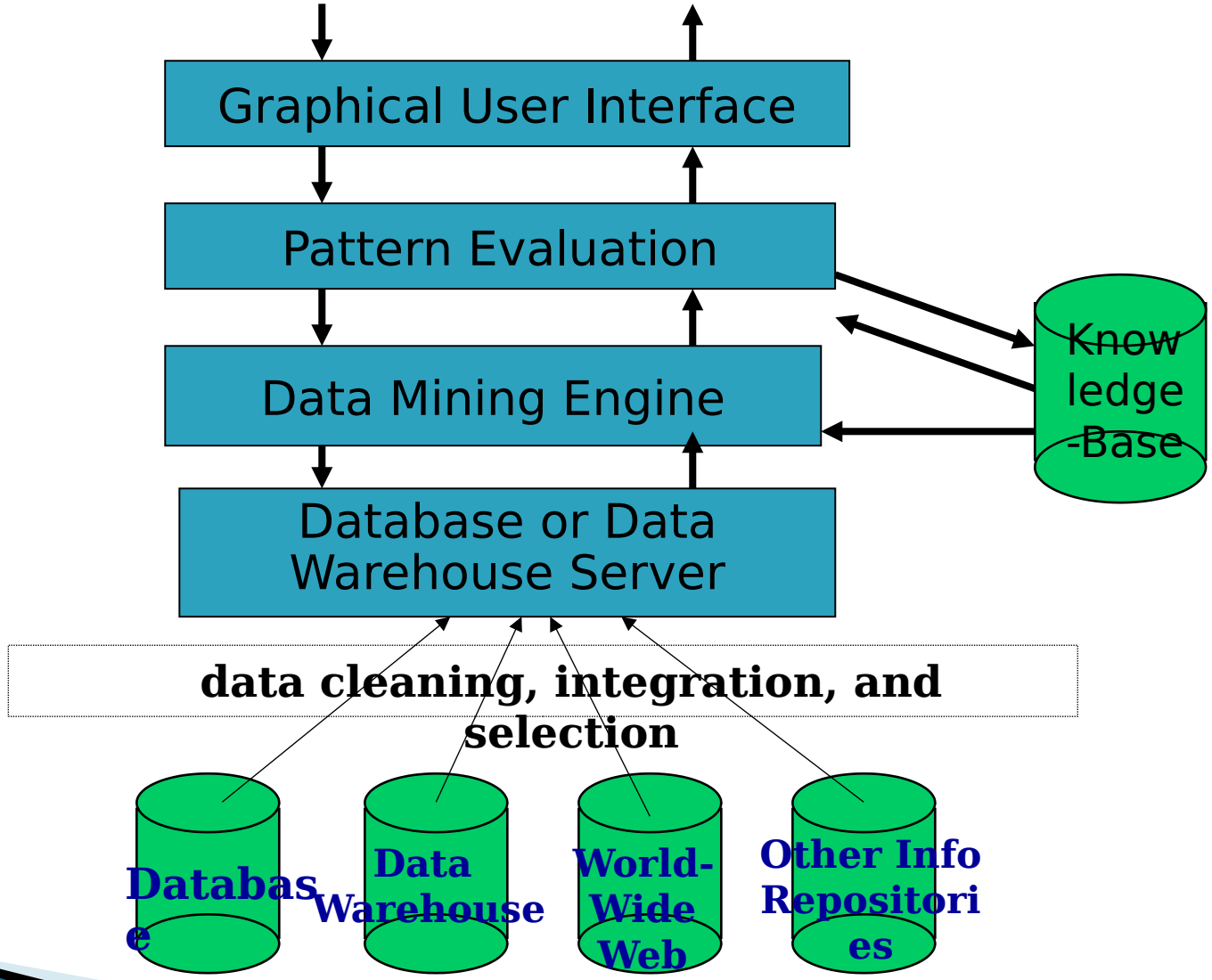
Conferences and Journals on Data Mining

- KDD Conferences
  - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
  - SIAM Data Mining Conf. (SDM)
  - (IEEE) Int. Conf. on Data Mining (ICDM)
  - Conf. on Principles and practices of Knowledge Discovery and Data Mining (PKDD)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
- Other related conferences
  - ACM SIGMOD
  - VLDB
  - (IEEE) ICDE
  - WWW, SIGIR
  - ICML, CVPR, NIPS
- Journals
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD

Data Mining and Business Intelligence



Architecture: Typical Data Mining System



Recommended Reference Books

- S. Chakrabarti. Mining the Web: Statistical Analysis of Hypertex and Semi-Structured Data. Morgan Kaufmann, 2002
- R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003
- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996
- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed., 2006
- D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, 2001
- B. Liu, Web Data Mining, Springer 2006.
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991
- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005